様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２０　日現在

機関番号：　３２６８９
研究種目：　若手研究(A)
研究期間：　2012～2015
課題番号：　２４６８００３１

研究課題名（和文）修飾・編集ＲＮＡの構造予測手法の研究開発

研究課題名（英文）Research on structure predictions of RNA with modified nucleotieds

研究代表者
浜田　道昭（Hamada, Michiaki）

早稲田大学・理工学術院・准教授

研究者番号：００５９６５３８

交付決定額（研究期間全体）：（直接経費）　　11,000,000円

研究成果の概要（和文）：修飾／編集塩基を含むRNAの構造予測に向けた情報技術の研究開発を行った．修飾／編集塩基を含む既知のRNAの構造データは極めて限られているため，このような限られたデータを用いて，効果的に構造予測を行う手法の開発を行う．具体的には，少数の２次構造データから２次構造の確率モデルを学習するための，半教師有り学習の方法を新しく開発を行った．また，RNAの統合Webサーバ Rtoolsを開発し，一般に公開した．

研究成果の概要（英文）：We have developed bioinformatic methods for predicting secondary structures including modified bases. Due to the limitation of the known structures with modified bases, we employed a semi-supervised learning approach for predicting RNA secondary structures using RNA sequences with and without secondary structures. Moreover, we have developed an integrated web server, Rtools, for performing various analyses based on RNA secondary structures.

研究分野：　バイオインフォマティクス

キーワード：　RNA　２次構造予測　修飾塩基　学習モデル　確率モデル　半教師有り学習

１．研究開始当初の背景

研究代表者の浜田は高精度な２次構造予測ツール CentroidFold に代表される２次構造を基盤とした RNA の情報解析技術を多数考案してきた．一方，修飾／編集 RNA は生物学的に様々な機能を有し，かつ，RNA 創薬でも利用される重要な研究対象であるにもかかわらず，修飾／編集 RNA の２次構造が予測可能なツールが存在しない．

２．研究の目的

修飾／編集塩基を含む RNA の構造予測に向けた情報技術の研究開発を行う．ただし，修飾／編集塩基を含む既知の RNA の構造データは極めて限られているため，このような限られたデータを用いて，効果的に構造予測を行う手法の開発を行う．具体的には，少数の２次構造データから２次構造の確率モデルを学習するための，半教師有り学習の方法を新しく開発を行う．

３．研究の方法

RNA の構造予測のためには，RNA の構造の確率モデルを利用することが広く行われている．修飾／編集塩基を含む RNA の２次構造の確率モデルを構築することが重要となるが，修飾／編集塩基を含む既知構造は限られているため，構造が未知の配列も利用した確率モデルの構築方法を新しく開発する．

４．研究成果

(1) 半教師有学習を用いた RNA の２次構造予測手法の研究開発と評価

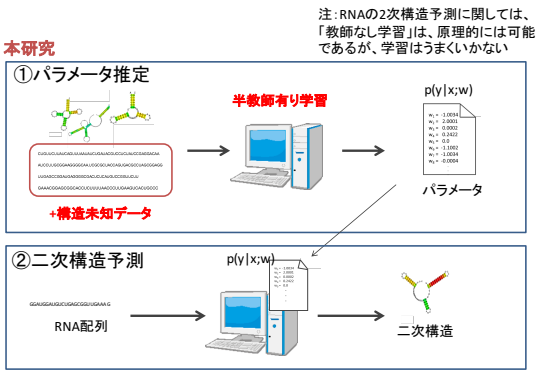

図 1 半教師有学習による２次構造予測

① 概要：RNA の２次構造の確率モデルを機械学習的な方法を用いて学習する場合には，従来は２次構造既知の RNA 配列が多数学習データとして必要であった．しかしながら，修飾塩基を含む構造データは極めて限られているため，２次構造既知と未知のデータを共に学習データとして利用可能とするための半教師有学習（semi-supervised learning）の方法を新しく開発した．概要は図１を参照して頂きたい．

② 提案モデル：提案モデルは，自然言語分野で利用されているモデル（Suzuki et al. ACL2007）を RNA の構造予測のモデルに拡張したものであり，生成モデル（generative model）である確率文脈自由文法（Stochastic Context Free Grammar: SCFG）と識別モデル（discriminative model）である条件付き確率場（Conditional Random Field; CRF）を組み合わせたハイブリッドモデル（Hybrid model）である（図 2）．最終的に得られるハイブリッドモデルは，識別モデルの一種となることがわかる．また，簡単な式変形により，CRF の特徴量に，SCFG による同時確率の情報が特徴量として追加したモデルとなることも示せる．

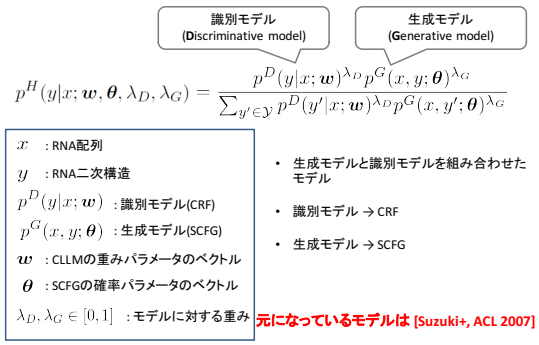

図 2 提案モデル（ハイブリッドモデル）

③ パラメタの推定：上記のハイブリッドモデルでは，識別モデル CRF のパラメタ w，生成モデル SCFG のパラメタ θ，２つのモデルに対する重みパラメタ λ の３種類のパラメタが存在する．これらのパラメタを，構造有りデータと構造なしデータから推定することが必要となる．本研究におけるパラメタの推定方法の概略を図３に示した．構造未知のデータは識別モデルのパラメタの推定の際に用いる．その他のパラメタは構造既知のデータを用いて行う．

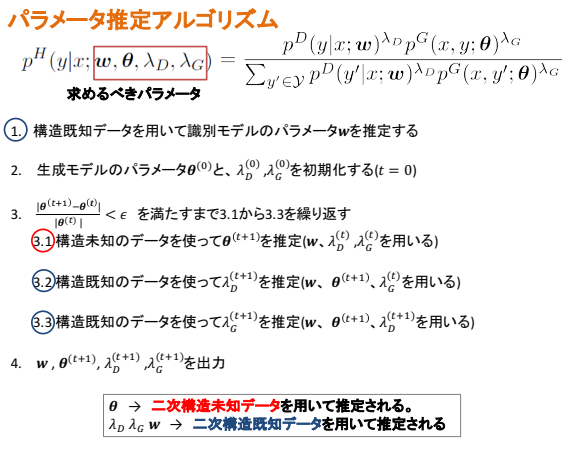

図 3 パラメタ推定の概要

④計算機実験による評価：Rivas らが用いたデータセット[Rivas et al, RNA 2012]を用いて提案手法の評価を行った．データセットの詳細は図に示した．この内 TrainSetA は２次構既知のトレーニングデータ，TrainSetB は２次構造未知のトレーニングデータとして用いた．

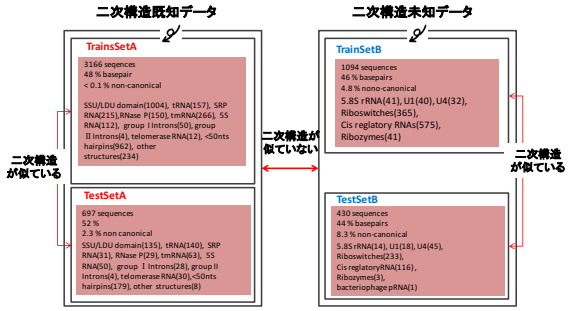

図 4 学習・評価に用いたデータセット

次に，提案（ハイブリッド），識別，生成モデルの間の予測精度の比較を行った．構造予測結果を図 5，6 に示した．これらの図より，特に TestSetB（似た構造のデータがパラメタ学習の際には利用していない）に対して，提案手法の既存モデル（識別モデルおよび生成モデル）に対する優位性が示されていることがわかる．すなわち，提案手法は構造未知の RNA の配列データを効果的にパラメタ推定に利用できていることがわかる．

Grammar	TestSetA			TestSetB		
	hybrid	disc	gen	hybrid	disc	gen
G6	**0.500**	0.488	0.478	**0.499**	0.476	0.462
G6s	0.502	**0.507**	0.489	**0.504**	0.496	0.475
basic.grammar	**0.573**	0.568	0.567	**0.574**	0.564	0.536

図 5 MCC による精度評価

さらに，TrainSetB における riboswitch ファミリーの有無と予測性能に関する評価を行った（図 6）．この実験においては，生成モデルのパラメタ推定に riboswitch ファミリーを含む場合と含まない場合の比較を行っている．

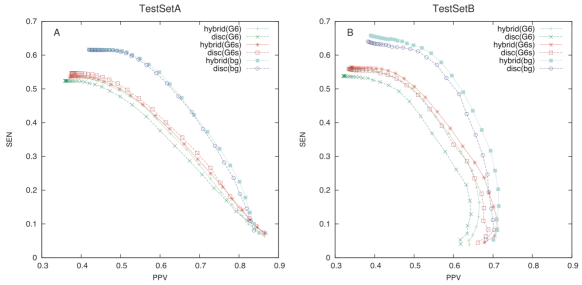

図 6 Sensitivity-PPV 曲線による精度評価

⑤結論：半教師有り学習手法を RNA の２次構造の学習に適用したハイブリッドモデルを開発した結果，提案モデルが既存のも出る（識別モデル，生成モデル）よりも優れた構造予測性能を示すことがわかった．したがって，既知構造データの少ない修飾塩基の２次構造予測を行う際にも利用可能であることが予想される．

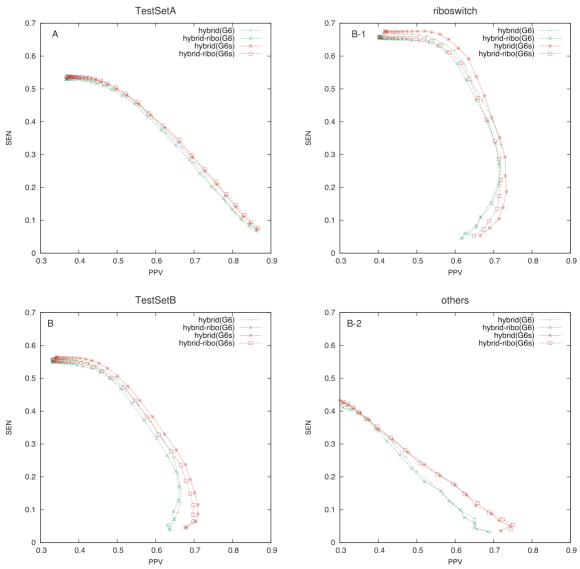

図 7 riboswitch ファミリを除いた場合の性能評価

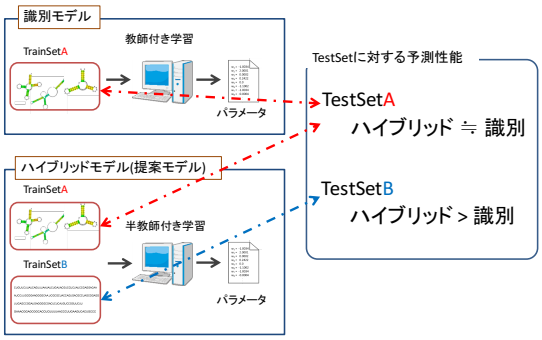

図 8 計算機実験のまとめ

(2) 統合 Web サーバ Rtools の開発：

RNA 情報解析ツールの普及のために，統合 Web サーバシステム Rtools（http://rtools.cbrc.jp/）の開発を行った．Rtools は，入力として RNA 配列を受け取り，構造予測に基づいた様々な解析を行う Web サーバである（図 9）．本成果は，国際学術雑誌 Nucleic Acids Research の Web Server Issue に掲載が決定している．

(3) 実験情報を利用した RNA の２次構造予測手法の開発

DMS-seq のデータを網羅的に収集をした．
細胞内の全転写産物（RNA）の総体はトランスクリプトームと呼ばれ、その全体像の把握は、生命のメカニズムを理解する上できわめて重要である．RNA は特異的な二次構造を形成して機能するため，二次構造の解明が肝　要となる．しかし，細胞内での RNA 二次構造

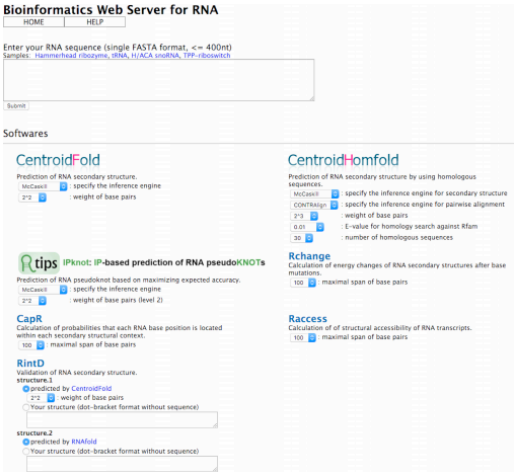

図 9　Rtools ウェブサーバ

は塩基 配列だけで決まるのではなく，様々な要素も組み合わさって決まると考えられており、その 予測は容易ではない．細胞内での RNA 二次構造の解明に向けて，近年，次世代シーケンサー技術を利用した RNA 二次構造予測の手法(以下，RNA 二次構造プロービング)が次々と開発されており，その解析結果から多くの知見が得られている．大きな期待が寄せられる一方で，高い疑陽性率や，低 発 現 転写物に対する予測が困難であるなどの問題点も指摘されており，標準的な解析手 法 はまだ確立されていない．そこで，本件では，RNA 二次構造プロービングの現状を調査し、既存の解析手法と解析パイプライン，および公開されている RNA 二次構造プロービングデータを収集し，また，既存の解析手法と解析パイプラインの性 能 評 価 を行うために，代表的な文献での解析の再現を行った．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計８件）

*: corresponding author(s), #: joint 1st authors

① Michiaki Hamada, Yukiteru Ono, Hisanori Kiryu, Kengo Sato, Yuki Kato, Tsukasa Fukunaga, Ryota Mori, Kiyoshi Asai, Rtools: a web server for various secondary structural analyses on single RNA sequences, Nucleic Acids Res.,［査読有り］
② Goro Terai#, Junichi Iwakiri#, Tomoshi Kameda, Michiaki Hamada*, Kiyoshi Asai*, Comprehensive prediction of lncRNA-RNA interactions in human transcriptome, BMC Genomics (2016) 17(Suppl 1):article no. 12.［査読有り］
③ Junichi Iwakiri#, Michiaki Hamada#, Kiyoshi Asai*, Bioinformatics tools for lncRNA research, Biochim Biophys Acta. 2016 Jan;1859(1):23-30.［査読有り］
④ Haruka Yonemoto, Kiyoshi Asai, Michiaki Hamada*, A semi-supervised learning approach for RNA secondary structure prediction, Computational Biology and Chemistry (2015) 57: 72-79.［査読有り］
⑤ Ryota Mori*, Michiaki Hamada, Kiyoshi Asai, Efficient calculation of exact probability distributions of integer features on RNA secondary structures, BMC Genomics (2014) 15(Suppl 10):S6.［査読有り］
⑥ Michiaki Hamada*, Fighting against uncertainty: An essential issue in bioinformatics, Briefings in Bionformatics (2014) 15 (5): 748-767.［査読有り］
⑦ Junichi Iwakiri*, Tomoshi Kameda, Kiyoshi Asai, Michiaki Hamada*, Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions, Bioinformatics (2013) 29 (20): 2524-2528.［査読有り］
⑧ Haruka Yonemoto, Kiyoshi Asai, Michiaki Hamada*, CentroidAlign-Web: a fast and accurate multiple aligner for long non-coding RNAs, Int. J. Mol. Sci. (2013) 14(3), 6144-6156; doi:10.3390/ijms14036144 (special issue: Non-Coding RNAs 2012).［査読有り］

〔学会発表〕（計４件）

① Michiaki Hamada, Comprehensive Prediction of lncRNA-RNA Interactions in Human Transcriptome, The Fourteenth Asia Pacific Bioinformatics Conference (APBC2016), San Francisco Bay Area, United States, Jan 11th-13th, 2016.
② Haruka Yonemoto, Kiyoshi Asai, Michiaki Hamada, A semi-supervised learning approach for RNA secondary structure prediction, The thirteenth Asia Pacific Bioinformatics Conference (APBC2015), HsinChu, Taiwan, Jan 21th-23th, 2015.
③ 浅井潔、浜田道昭、小野幸輝、RNA２次構造情報解析のための統合ウェブ、第 16 回日本 RNA 学会、愛知県、ウィンク愛知、2014 年 7 月 23 日（水）～25 日（金）
④ 浜田道昭、Centroid series: fundamental programs of sequence analysis for non-coding RNAs、バイオインフォマティクスとゲノム医療―その課題と将来展望―、2013 年 11 月 20 日(水) かずさ DNA 研究

所/産総研 生命情報工学研究センター共催ワークショップ. 東京 産総研.

〔図書〕(計　1　件)

① 瀬々潤, 浜田 道昭, 生命情報処理における機械学習：多重検定と推定量設計, 講談社 (2015)

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等：特になし.

６．研究組織

(1)研究代表者
浜田 道昭 (Hamada, Michiaki)
早稲田大学・理工学術院・准教授
研究者番号：00596538